DOCUMENT RESUME

ED 281 913 UD 025 451

AUTHOR Melville, S. Donald; And Others

TITLE Current Issues in Testing, Measurement, and Evaluation.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jan 87
CONTRACT NIE-P-85-0008
NOTE 17p.; In: Trends and Issues in Education, 1986 (see UD 025 435).
PUB TYPE Information Analyses - ERIC Information Analysis Products (071)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accountability; Achievement Rating; *Computer Assisted Testing; Educational Improvement; Educational Testing; Educational Trends; Elementary Secondary Education; Higher Education; *Minimum Competency Testing; Psychometrics; *Student Evaluation; *Teacher Certification; Test Construction; *Testing; Test Items
IDENTIFIERS Educational Issues; *Higher Order Learning; National Teacher Examinations; Teacher Certification Tests

ABSTRACT

In this report four educators discuss the issues which they see to be most current in the fields of testing, measurement and evaluation. The first section discusses the mastery of basic skills, defined by minimum levels of competence. Factors such as accountability, social policy, instructional implications, and psychometric issues are brought to bear on the subject. The second section examines problems more complex than those involved with assessing basic skills. Identifying and defining higher order skills, designing a sound curriculum, and deciding whether available instruments are adequate for assessing higher order skills are among the goals which must be achieved to adequately teach and test higher order skills. The third section points out some concerns related to testing teachers before they begin to practice their profession. Two major trends, using the National Teacher Examinations from the Educational Testing Service and using state programs to develop teacher certification tests, are presented. A state-of-the-art survey, detailed in section four, describes computer-assisted educational testing as it is used for writing test items, constructing tests, administering tests, scoring and analyzing results, and record keeping. A list of references is included. (PS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281913

Current Issues in Testing, Measurement, and Evaluation

By
S. Donald Melville, Jacob G. Beard, C. Philip Kearney,
Rodney Roth, and Jason Millman

Chapter 15 of
Trends and Issues in Education, 1986

Erwin Flaxman
General Editor

Prepared by
Council of ERIC Directors
Educational Resources Information Center (ERIC)
Office of Educational Research and Improvement
U. S. Department of Education
Washington, D. C. 20208

January 1987

UD025451

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

413-89

CURRENT ISSUES IN TESTING, MEASUREMENT, AND EVALUATION

S. Donald Melville, Director, ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service, Princeton, NJ; Jacob G. Beard, Florida State University; C. Philip Kearney, University of Michigan; Rodney Roth, University of Alabama; Jason Millman, Cornell University

Four educators describe the issues which they see to be most current in the fields of testing, measurement and evaluation. The mastery of basic skills, defined by minimum levels of competence, is discussed by Jacob G. Beard, in "Minimum Competency Testing." Issues such as accountability, social policy, instructional implications, and psychometric issues are brought to bear on the subject. C. Philip Kearney, in "Assessment of Higher Order Skills," examines a set of problems more complex than those involved with assesing basic skills. A clear definition of what constitutes higher order skills, a sound curriculum design, and available instruments for assessing higher order skills are among the goals which must be achieved to adequately teach and test higher order skills. In "Testing Teachers for Initial Certification," Rodney Roth points out some of the concerns related to testing teachers before they begin to practice their profession. Two major trends, using the National Teacher Examinations from Educational Testing Service and using state programs to develop teacher certification tests, are presented. A state-of-the-art survey by Jason Millman, "Educational Testing and the Computer," describes computer-assisted educational testing as it is used for writing test items, constructing tests, administering tests, scoring and analyzing results, and record keeping.

Minimum Competency Testing

During the last decade many school systems began defining minimum levels of competence for their students and constructing tests to measure whether the students had achieved these minimums. These minimum competencies usually included the basic skills of reading, writing, and arithmetic and their application. The term "minimum competency testing" acquired special meaning from this activity. Considerable controversy arose when, in 1976, Florida passed a law ~~which~~ requiring high school students to pass a minimum competency test in order to graduate. A class-action lawsuit was brought against Florida's school system in an effort to block the use of the test as a graduation requirement. The courts upheld the rights of school systems to establish minimum standards of competency for graduation, and many other states now have similar laws. The controversy has continued and is focused on the following issues.

Accountability

During the 1970's there was considerable criticism of the schools and accusations of lowered achievement. To many, minimum competency testing was

seen as a means of holding the schools accountable for graduation of literate students who could perform the basic skills of reading, writing, and arithmetic. All students would be tested for minimum competencies and failures would be remedied before graduation. Students who were unable to remedy their weaknesses and pass the test before graduation would be given certificates, but not high school diplomas.

Many educators have expressed concern about the effects of minimum competency programs on the overall school curriculum, and the level of achievement resulting from the programs. There is speculation that the minimum will become the maximum competencies at the expense of higher learning levels. Such an effect has not been demonstrated; however, some political and educational leaders have responded to the concerns by adding testing programs measuring higher levels of achievement.

Statewide minimum competency testing is inconsistent with the concept of local control. Some freedom of districts to determine what is taught in the schools must be relinquished to the state when state testing programs are established. However, the curriculum for most schools is already rather fully determined by state and national policies. The idea of each school district's separately determining a unique curriculum is not consistent with current practice.

Social Issues

Minimum competency testing is seen by some as social policy. Cohen and Haney (1980) argued that it was another in a long line of educational minimums which began when elementary education was made compulsory and was followed periodically by increasing requirements for formal education. Previous minimums have been phrased in terms of age or years of schooling. Cohen and Haney point out that while the establishment of official minimums has the appearance of equalizing achievement, history shows that it merely initiates a new competition for superiority.

Minimum competency testing has also been characterized by its opponents as a racist means of denying educational credentials such as high school diplomas to minority, and particularly black, students. This argument is based on the historically greater failure rate of black than of white students on these and other academic achievement tests. Proponents of minimum competency testing argue that it is a means of improving the achievement of marginal students by identifying achievement deficiencies and ensuring that all students receive a basic education.

Instructional Implications

Minimum Competency Testing programs must be coordinated with the instructional program. The tests must have both curricular and instructional validity. That is, they must measure instructional objectives which are included in the established curriculum and which are actually taught to all of the students.

Remedial instruction should be made available to students who fail the test before retaking it. This usually requires additional funding to ensure that adequate remediation is given.

A basic premise of educational systems which adopt minimum competency testing is that credit should be given for accomplishing instructional objectives rather than for spending time in programs. This idea leads naturally to the implementation of various instructional design concepts such as: diagnosis and prescriptive learning, individualized instruction, and optimally designed instructional materials. These concepts have been introduced before, but have had limited success in achieving widespread or long-term implementation. However, effective minimum competency testing virtually necessitates the use of such systems.

Psychometric Issues

When minimum competency tests are used to make decisions having serious consequences for students, the psychometric properties of the test scores become especially important. Individuals denied high school diplomas on the basis of minimum competency test results have sued the educational system. They have charged that the use of inadequate tests constituted violation of the due process and equal protection clauses of the Fourteenth Amendment to the Constitution. Therefore, users of such test results should make sure that the testing program conforms to the standards of quality set forth by the testing profession. This includes adherence to the Standards for Educational and Psychological Tests published jointly by the American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). The following criteria are especially important for minimum competency tests.

- The tests must have content, curricular, and instructional validity; that is, they must test material which has been taught to all the students.
- Students must be given adequate warning of new standards for graduation.
- The test scores which assign students to the categories of pass or fail must be reliable for that purpose.
- The passing score representing the achievement of minimum competency must be arrived at rationally and the level of skill it represents must not fluctuate from one test administration to another.

- o The test must not contain items which are biased for or against any racial, ethnic, religious, sex, or other group through characteristics other than the measurement of stated instructional objectives.

- o Absolute security of the tests must be maintained.

- o Test administrations must be standard at all testing sites.

Trends

Minimum competency testing continues to be used as a requirement for high school graduation and has been introduced at other levels of education. For example, several states have installed tests of minimum competency for college sophomores and for teacher certification.

Several states have responded to concerns about lowered achievement expectations by initiating testing programs which measure levels of achievement beyond the basic skills within, or in addition to, their minimum competency testing programs.

Assessment of Higher Order Skills

The teaching and testing of higher order skills is fast taking on the characteristics of a nationwide educational reform movement. Several states are developing and implementing programs aimed at assessing higher order skills, local school districts are moving rapidly to adopt curricular programs that emphasize the teaching of higher order skills, educational textbook publishers and testing companies are becoming increasingly active in this area, and conferences, symposia, and workshops on this topic are springing up all across the land.

The growing concern over higher order skills stems principally from a recognition that the nation's pupils, while demonstrating modest improvement in the basic skills, are falling far short of achieving mastery of thinking skills--long considered one of the major instructional goals of schooling. There is ample evidence to support this contention--a decline in SAT and ACT scores over the past several years, results from the National Assessment of Educational Progress demonstrating a lack of analytical skills among the nation's pupils, and results from state testing programs suggesting shortcomings (Harnischfeger & Wiley, 1975; NAEP, 1981; Baron, 1985).

The higher order skills are increasingly becoming a principal focus of state level assessment efforts, a phenomenon which bodes well for those advocating a strong curricular emphasis on the higher order skills--for tests drive the curriculum, particularly state tests. What the state tests determines, in large part, what the schools teach and the relative degree of emphasis placed on the subjects and areas tested in relation to other subjects and areas of the curriculum (Rudman, 1985).

However, the assessment of higher order skills--whether at local, state, or national levels--poses problems that are more complex and substantially different from those posed by the assessment of basic skills and other subjects traditionally found in the school curriculum. The first of these problems centers on the lack of clear definition of what constitute higher order skills. What precisely is it we are talking about when we use the term "higher order skills"? A second problem is whether we are better advised to teach--and test--higher order skills as a separate subject in the curriculum, divorced from particular content areas such as reading, mathematics, and science; or whether we are better advised to teach and test higher order skills as an integral part of one or more subject areas. A third problem focuses on the availability, or unavailability, of instruments to assess student attainment in the higher order skills. Is there a need for considerable test development work or are valid and reliable measures already in existence? And there are other problems--for example, questions of a "one-tiered" versus a "two-tiered" approach (mastery of basic skills, then mastery of higher skills). Still other problems: the costs and benefits of using writing samples in measuring these skills, and questions of every-pupil testing versus a sampling of pupils.

The problem of lack of clear definition is particularly acute. "Higher order skills" is one term used to describe thinking skills. Other terms abound--critical thinking, higher order thinking skills, higher level skills, reasoning, intelligence, creative thinking, lateral thinking, informal logic, to name a few. The problem is not only to decide among these names but, perhaps more importantly, to choose what definition or conception of thinking will guide teaching and testing activities. At the present time, there seems to be little if any consensus on names or definitions. For the parent, the answer is easy: "What I want is for you to teach my child to think." For the profession, the answer is much more complex. It includes such notions as a habit of reflective thinking; a disposition or willingness to think critically, assertively, and habitually; more difficult subject matter content; critical reasoning skills; skills that go beyond recall or learning of facts; and a literal laundry list of other cognitive activities (Beyer, 1983; Kean, 1985). One acknowledged leader in the field chooses the term "critical thinking" and defines the concept as "reasonable reflective thinking that is focused on deciding what to believe or do" (Ennis, 1985). Another defines "thinking" as "the operating skill with which intelligence acts upon experience" (de Bono, 1983). Still another offers a definition of "higher order thinking skills" as:

> those skills that go beyond straight recall or learning of facts....problem identification and problem solving, evaluation of information and of arguments, deduction, inference, taking alternate points of view, creating reasonable arguments in support of a position, and making decisions. (Fremer & Daniel, 1985)

Thus, when it comes to defining precisely what thinking skills mean, it seems there is no consensus but great diversity in both terms and concepts. For

those who would include higher order skills in a state assessment program, the first task is one of settling on a meaningful and useful definition.

The second problem, whether the higher skills should be taught and tested as a separate subject area or embedded or infused into existing subject matter and tested in like fashion, also lacks resolution, even though most people favor the latter. Still, the former approach, teaching and testing thinking skills as a separate topic area, has strong support among several leaders in the field. Sternberg, for example, argues that the better strategy is one that assumes intervention at the level of mental processes, and that pupils can be taught when and how to use particular mental processes, and how to combine those processes into strategies that lead to problem solutions (Sternberg, 1984). He argues for three programs to teach the components of intelligence--intelligence being his choice of name and definition of higher order skills. The three are Feuerstein's "Instrumental Environment," Lipman's "Philosophy for Children," and "The Chicago Mastery Learning Program" (Sternberg, 1984). Another acknowledged leader in the field, Edward de Bono, also argues for the direct teaching of thinking as a skill; he calls for setting aside a place in the school program so that pupils, teachers, and parents will recognize that thinking skills are taught directly (de Bono, 1983). However, de Bono is much less sanguine about ability to assess thinking. He argues that our present measures are not up to the job because they do not observe the thinker's composite performance. A third acknowledged leader, Robert Ennis, supports the inclusion of critical thinking as an inherent part of traditional subject matter, even though some contend that he favors both approaches (Ennis, 1985; Baron 1985). While there is ample evidence that either approach can work, most research seems to support Ennis's view--namely, that instruction in thinking skills should be present across subject areas and throughout the grades (Beyer, 1983; ETS, 1984; Fremer & Daniel, 1985; Kean, 1985).

Still, Connecticut, in its state level assessment programs, is using both approaches apparently with equal success. It systematically integrates higher order thinking skills into its assessment of the subject matter domains covered in the ongoing Connecticut Assessment of Educational Progress while, at the same time, it explores a variety of additional formats to measure critical thinking and reasoning skills separately and more directly in its newly developed Mastery Testing Program (Baron, 1985). Michigan, on the other hand, is moving to test thinking skills as part of a revised every-pupil reading and math assessment to be administered at grades 4, 7, and 10 and as a newly developed every-pupil writing assessment at grades 5, 8, and 11 (Michigan Department of Education, 1986). In Florida, the emphasis also is on testing higher order skills within content areas (Fremer & Daniel, 1985). Thus, while we see both approaches pursued in the assessment of higher order skills, current practice seems to give an edge to teaching and testing such skills as embedded parts of traditional subject areas.

The third problem, whether instruments currently available are adequate for assessing higher order skills, also admits of different responses. Some argue that commercially available standardized achievement tests include items that measure higher order skills, and that scores and sub-scores from these instruments can provide useful and valid information on pupil attainments of higher order skills (Fremer & Daniel, 1985; Kean, 1985). Others contend there are no topic-specific critical thinking tests available, but only tests which attempt to cover critical thinking as a whole, or focus on one aspect of critical thinking (Ennis, 1985). Still others--particularly those who develop and implement state level assessment programs--argue that, while much developmental work remains, there are measures of higher order skills that can be incorporated into ongoing programs, so state level efforts need not wait on long-term developmental efforts (Baron, 1985; MDE, 1986).

There are other problems. Should there be a two-tiered approach? Should higher-order skills be assessed only after a pupil has demonstrated mastery of the basic skills? Should writing samples be used to assess higher order skills? If so, what form should these take and how should they be scored? Is it important to test every pupil at every grade level? Or can the state accomplish its purposes by sampling grades and sampling pupils? While research can be helpful in addressing problems of these types, their ultimate resolution may depend more on the policy values and policy culture prevailing in any particular state.

Testing Teachers for Initial Certification

Testing teachers before they begin to practice their profession is not a recent phenomenon. The first official endorsement of teacher testing occurred in the colonial era (Vold, 1985). The General Assembly of Virginia in 1686 requested that every county appoint a person who would examine and license schoolmasters. The testing of teachers for county certification was dominant throughout the United States from 1860 until the early 20th century.

The development of normal schools to train teachers and the approval of teacher training programs by state departments of education led to an elimination of testing teachers for certification by the 1920s. The American Council on Education did, however, establish the National Teacher Examination in 1940. Initially, it was used by local school districts to help with teacher selection; only recently has it been used for certification.

The testing of teachers for certification has resurfaced in the past decade; a majority of states currently test teachers for certification and more states plan to start. The rebirth resulted from several major factors. Two of these factors were declining test scores and an oversupply of teachers. Another was the large scale press coverage given to a very few letters written by teachers to parents. The letters contained errors in grammar and spelling.

The rest of this section will present two major trends and procedures in the testing of teachers for initial certification and briefly discuss some current problems or dilemmas facing policy makers, researchers, and persons involved with teacher testing.

Major Trends

One trend is to use the National Teachers Examinations (NTE) from Educational Testing Service (ETS). The use of this test can be traced, in part, to two court decisions from the Carolinas. South Carolina started using the NTE to assign different grades of teacher certificates shortly after it was developed. The type of certificate affected salaries and salary increases.

In 1971, ETS issued guidelines stating that passing scores or cut-scores should be based on validation studies. In 1975, a District Court in North Carolina issued a decision requiring objective proof by the State of North Carolina of the relationship between the minimum score requirements on the NTE and the State's objective of certifying teachers who were at least minimally competent. Based on this decision, South Carolina authorized an NTE validity study by ETS.

The validity study conducted by ETS assessed the extent to which the content of the NTE tests represents the content of the teacher training programs. Teacher educators were asked to make several judgments about the overall test specifications and teacher training programs. They were further asked to review each question on the test and judge its appropriatenes. A question was considered "content appropriate" if at least 51% of the judges indicated that at least 90% of the students would have had an opportunity to learn the content.

The cut-scores derived from the validation study and adopted by South Carolina for initial teacher certification were challenged in court. In January, 1978, the United States Supreme Court announced that it had affirmed the April, 1977, decision of a Federal District Court upholding South Carolina's use of the NTE for certification. This decision prompted several other states to adopt the NTE with cut-scores based on similar validation procedures.

The United States government issued the Uniform Guidelines on Employee Selection Procedures just after the Supreme Court decision on the NTE use in South Carolina. These Guidelines apply to tests used for hiring, promotion, and licensing and certification to the extent that licensing and certification may be covered by Federal equal employment law. These Guidelines require that tests be validated in terms of job relatedness. This prompted Roth (1982) to develop a new validation procedure for his NTE study for the state of Arkansas.

This NTE study used teachers and teacher educators to judge each test item. The judges rated the relevance of the content measured by each question against the domain of knowledge they believed essential for a minimally qualified entry-level person. Most NTE validity studies done since 1982 have assessed both job relevance and the relationship to teacher training programs.

Another current trend is for states to develop their own teacher certification tests. In practice, this typically means that states contract with the National Evaluation Systems (NES) for test development and subsequent scoring and reporting services. Georgia was the first state to develop its own tests for teacher certification. Interestingly, Georgia decided not to use the NTE. This was based in part on a court decision concerning its use of the NTE for awarding an advanced teacher certificate. Georgia had selected an NTE cut-score that was not based on a validity study for the certificate. In January, 1976, a District Court ruled that the test had no rational relationship to the purpose of the certificate. The Court also indicated that a state must show a valid relationship between a general national examination and the specific duties performed by a teacher in the state.

States that develop their own tests typically use procedures following the Uniform Guidelines on Employee Selection Procedures. This means that the tests are designed based on the knowledge needed to teach a specific subject in the state. Elliott (1986) presents various procedures used by several states to develop their own tests. The key component in these procedures is a job analysis. It includes some determination of the critical and frequently performed elements of the job. The job analysis typically begins with a large number of content or topic objectives derived by content experts to define the scope of the teaching field. Teachers rate each objective according to its essentiality and the amount of time spent teaching the content. The results of this process determine the specific objectives for which test items are developed. The items are evaluated for their congruence with the objectives. The remaining items are field-tested in order to produce appropriate item and test statistics. These results are used to produce the final or actual certification test.

Problems or Dilemmas

At the outset, a major dilemma faces policy makers who must choose whether to use the NTE or develop their own test. Some of the advantages of the NTE are that the test is available; it is administered by a large and creditable testing firm; it has been used for over 45 years; and its use was upheld by the Supreme Court. One disadvantage is that appropriate tests are not available for certain certification fields. In addition, state validation studies that use current validation guidelines might indicate that the NTE is not appropriate or that the derived cut-scores are extremely low.

The major advantages of state-developed tests are that the tests can be developed for each certification field and the tests cover the essential knowledge needed to teach a field in the State. The major disadvantages are the time and cost involved for test development. A potential problem is that state-developed procedures have not been tested in the courts.

A second problem for policy makers concerns what to test. Some states test the content in the certification field; other states test professional knowledge; and still others test general knowledge. The professional and legal guidelines for employment testing seem to indicate that the further

one moves away from the specific content needed for the position, the more difficult it is to show job relatedness. For example, potential math teachers should have literature as part of their training program. Should they, however, be tested on literature as well as math in order to be certified to teach math?

A major problem for educational researchers and people who develop state tests or validate existing tests is to determine what guidelines and standards are appropriate. The Supreme Court decision for South Carolina indicates that a validity study based on the teacher training program is appropriate. The Uniform Guidelines would seem to indicate that the South Carolina procedure was not appropriate. Rebell (1986) states the problem by saying that regarding the law, there is an unresolved technical issue whether Title VII and the Equal Employment Opportunity Commission (EEOC) Guidelines apply to licensing or credentialing examinations. He also raises a question of precisely how those validation standards, that were created largely in the context of individual employer job selection tests, should be implemented in the conceptually distinct licensing or credentialing context. The 1985 Standards for Educational and Psychological Tests (American Psychological Association) have also added a section on professional and occupational license and certification. These standards seem to indicate similar procedures found in the Uniform Guidelines. The impact of the Debra P case in Florida on certification testing is another unknown variable. It reintroduces the question of curricular and/or instructional validity.

After the validation guidelines or test development procedures have been decided, a new series of decisions has to be made. These concern professional judgments that have to be thought out during the process. Some examples are: Should the percentage who typically answer an item correctly be provided for the judges who are making item probability estimates; what is an appropriate standard to judge item relevance, or item essentiality, or content coverage; and what roles should various standard errors have on the process.

Conclusion

Certification is intended to protect the public. Teachers, like most professions, should be tested for initial certification. The problems associated with the process are complex, but not unsolvable.

Solutions are needed because society can neither afford to have incompetent teachers teach our children, nor can it afford to deny competent persons the chance to practice their chosen profession.

Educational Testing and the Computer

Computers are involved in educational testing in five areas: (a) writing the test items, (b) constructing the tests, (c) administering the tests, (d) scoring the tests and analyzing and interpreting the results, and (e) keeping test records. This survey describes the state of the art with respect to computer-assisted educational testing.

Writing the Test Items

Of the five areas, the writing of items has been least influenced by computers. Thus far, the potential of the computer to compose item content has not been realized.

The first attempt at computer-generated item writing took place in 1968 when two educational researchers, H.G. Osburn and David Shoemaker, working under a U.S. Office of Education grant, developed a scheme by which the computer would construct questions about statistics. This scheme worked by completing a fixed part of the question called an item shell with words or numbers randomly selected from a set of possibilities called a replacement set. For example, a true-false question might be generated by the computer by putting together the shell, "The middle number in a distribution is called the" and a randomly selected word from the replacement set, "mean, median, mode." Note that in this simple example three variations of the true-false question are possible.

In item shell and replacement set schemes, every word that appears in a test question is first thought of by the item writer and entered into the computer. The computer is relegated to the trivial task of picking the words or numbers and putting them together using straightforward algorithms to produce the test questions. Although some attempts to have the computer "think" like a test constructor have been carried out, for the present the computer provides scant practical help to the item writer.

Constructing the Tests

The computer is used extensively to build tests, especially by commercial publishers and governmental agencies. This application is made possible by collections of items called item banks. Occasionally, items are kept only on paper while documentation of each item--its statistical properties, content descriptions, and so forth--are fed into the computer. The computer then can pick a collection of items that meets the statistical and content specifications of the test builder. It is then left to the test constructor to assemble the test manually.

More common, however, is the situation in which the items themselves are entered into the computer, together with several pieces of documentation. When the items are stored, the computer can both select appropriate items and construct and print the test itself. The successful and extensive use of the computer to <u>assemble tests</u> is in contrast to its minimal use to <u>write items</u>.

Instructors who teach the same subjects may develop an item bank which they share. Sometimes they obtain the item bank from a state or local agency or from a commercial source; at other times they construct their own items, perhaps beginning by using items available from others. The Northwest Regional Educational Laboratory, 300 SW Sixth Avenue, Portland, Oregon 97204, provides listings of available item banks and reviews of existing microcomputer programs that will construct tests from item banks. Most of the programs are too limited to be very useful. A few of the more recent ones, however, show promise.

Millman and Arter (1984) provide detailed information about item banks and test construction. They describe a wide variety of item banks, outline their advantages and disadvantages, list the conditions under which item banks have the most potential value, and provide an extensive set of questions to be asked in designing item-banking systems.

Large-scale test development programs will become increasingly computerized. Individual teachers can expect to assemble their tests from computerized item banks as quality software and microcomputers become available.

Administering the Tests

The glamour area in educational testing these days is computer administration of tests. What makes this area so fascinating is the ability to program the computer to consider a student's prior answers when picking the next question; that is, to select items for administration based on the student's previous responses. Thus, the examination given to each student can be tailored or adapted to his or her level of ability. It is this adaptive, tailored, response-contingent feature that gives computer-administered testing its major advantage over conventional test administration.

Adaptive testing, as it is most frequently called, has been put to use to help solve three knotty testing problems. The first is getting more measurement precision with fewer test items. It is a fact of psychometric life that the more test items given to a student, the more accurately the student's level of achievement or ability can be assessed. But teachers and students alike object to tests that take a long time to complete. Because the level of difficulty of the items a student is given under adaptive testing corresponds to the student's level of performance, they carry maximum information about the student's ability, with the result that adaptively administered tests can provide the same degree of precision as traditionally administered tests while using about half as many items.

The second problem attacked by adaptive testing is that of making test items simulate tasks that the student might face on the job or in other out-of-school situations. In adaptive testing, the computer can be programmed to permit students to progress through a program situation and to provide students with appropriate feedback. For example, in patient-management problems, a medical case is presented and the medical student indicates what actions should be taken. These actions might include observing the patient's physical condition, ordering laboratory tests, or prescribing medication or other treatments. The result of each action is given to the student, who proceeds to answer additional questions about further treatment.

The third problem that adaptive testing is well suited to handle is diagnosis of student learning problems. When a student misses a test question, the computer can be programmed to administer carefully selected similar items that can pinpoint the student's misconceptions or gaps in knowledge. With such information, the teacher can provide appropriate remedial instruction.

Although some large testing programs have begun to administer tests by computer, with positive reactions from those examined, it will be some time before classroom teachers routinely give their tests by computer. Tests embedded in instructional computer software are the exception. Questions asked of learners are an integral part of the teaching material, and such testing is often so nonintrusive that the students are not aware they are tested.

Scoring the Tests and Analyzing and Interpreting the Results

For many years, groups who administered many objective tests scored their own answer sheets by hand. Now desk top scoring machines connected to a microcomputer are available for a price that enables local schools and small colleges to have their own automated scoring and test reporting system. In a few more years, a majority of the medium- and large-sized school districts may score and report objective tests using locally owned equipment.

Computers have also been used to score short-answer questions and to grade essays. The procedure typically consists of matching the student's answer to key words provided by the test constructor. If the student supplies the key words or acceptable variations, credit is given for the answer. Somewhat aside, it seems that the science of short-answer and essay test scoring has not made any noticeable progress in the last 10 or 15 years, nor is it likely to do so using present methods.

A traditional function of computers in testing has been to analyze item and test data. The prowess of computers to manipulate numbers has never been doubted, and computers continue to provide test developers with a much valued service in this regard. Using item data stored in item banks, some of the more sophisticated programs can predict the score distribution and other test results before a planned test is actually administered.

Computer interpretation of test results, particularly those of psychological tests, is the most controversial of all aspects of computer testing. Many computer companies now administer and <u>interpret</u> the results from interest, vocational, personality and intelligence tests. The controversy stems in large part from the secrecy that surrounds the algorithms the computer uses to produce various interpretations. How the computer decides that a job applicant is a good risk or that client has suicidal tendencies is often shrouded in proprietary secrecy, and the validity of these interpretations remains uncertain.

Keeping Test Records

Another task to which computers are well suited is keeping track of test performance. Computers can store results in a record or grade book, produce grade reports, and develop a profile of test results for an individual student or for the class as a whole. Microcomputer programs that perform

these functions are readily available and relatively inexpensive. The computer can be programmed to keep track of other statistics in addition to test scores: among these, the time taken to answer each question, the attractiveness of each foil in a multiple-choice item, and the proportion of students who answered each item correctly.

As discussed here, computers are employed in several areas of educational testing. The functions of computers in these areas can be integrated, which may lead to more efficient and acceptable testing practices. Using items from a bank, the computer can assemble and administer a test and, because the responses of computer-administered tests are entered directly into the computer, it can quickly score, record, and interpret the results. As computers and programs for carrying out these tasks become more readily available, we can expect a greater proportion of testing activities to be aided by computer. Although the computer can make the process easier to implement, the educational benefit that accrues to the student will depend on the quality of the items that make up the tests and on how the results are put to use.

References

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, DC: American Psychological Association.

Baron, J. B. (1985, February). Assessing higher order thinking skills in Connecticut: Some results, some lessons we've learned and the challenges ahead. Paper presented at the 25th annual Michigan School Testing Conference, Ann Arbor, MI.

Beard, J. G. (1979). Minimum competency testing: A proponent's view. Educational Horizons, 58(1), 9-13.

Berk, R. A. (1984). A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press.

Beyer, B. K. (1983, March). Improving thinking skills--Defining the problem. Phi Delta Kappan, 703-708.

Bureau of National Affairs. (1979). Uniform guidelines on employee selection procedures. Washington, DC: Author.

Cohen, D. K., & Haney, W. (1980). Minimum--competency testing and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), Competency testing: Motives, models, measures, and consequences (pp. 5-22). Berkeley: McCutchan.

de Bono, E. (1983, June). The direct teaching of thinking as a skill. Phi Delta Kappan, 703-708.

Elliott, S. M. (1986). Teacher certification testing technical challenges: Part II. In W. P. Gorth & M. L. Chernoff (Eds.), Testing for teacher certification (pp. 139-154). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Debra P. v. Turlington, 564 F. Supp. 177 (M. D. Fla. 1983).

Educational Testing Service. (1984). Critical thinking. Princeton, NJ: Author.

Ennis, R. H. (1985, October). A logical basis for measuring critical thinking skills. Educational Leadership, 43, 44-48.

Fremer, J., & Daniel, M. (1985, February). The assessment of higher order thinking skills. Paper presented at the 25th annual Michigan School Testing Conference, Ann Arbor, MI.

Harnischfeger, A., & Wiley, D. (1975). Achievement test score decline: Do we need to worry? Chicago: ML--Group For Policy Studies, CEMREL, Inc.

Jaeger, R. M., & Tittle, C. K. (Eds.). (1980). Minimum competency testing: Motives, models, measures, and consequences. Berkeley: McCutchan.

Kean, M. H. (1985, February). Assessing higher order thinking skills: First examine the foundation. Paper presented at the 25th annual Michigan School Testing Conference, Ann Arbor, MI.

Michigan Department of Education. (1986, February). Memorandum to the State Board of Education from Phillip Runkel.

Millman, J., & Arter, J. A. (1984). Issues in item banking. Journal of Educational Measurement, 21, 315-330.

National Assessment of Educational Progress. (1981). Reading, thinking, and writing: Results from the 1979-80 assessment of reading and literature. Denver: Education Commission of the States.

Osburn, H. G., & Shoemaker, D. M. (1968). Pilot project on computer generated test items. Washington, DC: Office of Education. (DHEW Grant 1-7-068533-3917)

Pipho, C. (Ed.). (1978). Minimum competency testing. Phi Delta Kappan, 59(9).

Rebell, M. A. (1986). Recent legal issues in competency testing for teachers. In W. P. Gorth & M. L. Chernoff (Eds.), Testing for teacher certification (pp. 59-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Roth, R. W. (1982). Validation of the NTE: Arkansas style. (Report No. TM 820 724). University, AL: University of Alabama, College of Education. (ERIC Document Reproduction Service No. ED 222 566)

Rudman, H. C. (1985, February). Testing beyond minimums. Paper presented at the 25th annual Michigan School Testing Conference, Ann Arbor MI.

Sternberg, R. J. (1984, September). How can we teach intelligence? Educational Leadership, 34-48.

Vold, D. J. (1985). The roots of teacher testing in America. Educational Measurement: Issues and Practice, 4(3), 5-7.